Analyse d'opinion et langage figuratif dans des tweets : présentation et résultats du Défi Fouille de Textes DEFT2017

Farah Benamara[1] Cyril Grouin[2] Jihen Karoui[1]
Véronique Moriceau[3] Isabelle Robba[4]
(1) IRIT, Université de Toulouse
(2) LIMSI, CNRS, Université Paris-Saclay
(3) LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay
(4) LIMSI, CNRS, UVSQ, Université Paris-Saclay
{farah.benamara, jihen.karoui}@irit.fr
{cyril.grouin, veronique.moriceau, isabelle.robba}@limsi.fr

RÉSUMÉ

La détection automatique du langage figuratif dans les réseaux sociaux est un sujet de recherche extrêmement actif principalement en raison de son importance pour améliorer les performances des systèmes d'analyse d'opinions. Pour la première fois, l'édition 2017 du Défi Fouille de Texte (DEFT) s'intéresse à l'influence du langage figuratif (en particulier l'ironie, le sarcasme et l'humour) dans l'analyse d'opinions à partir de tweets en français. Trois tâches de niveaux de complexité croissants ont été proposées aux participants : (T1) déterminer la polarité globale des tweets non figuratifs, (T2) déterminer si un tweet contient ou non du langage figuratif, et (T3) déterminer la polarité globale des tweets figuratifs et non figuratifs. Douze équipes ont participé à ce défi. Les meilleurs résultats, en macro f-mesure, sont de 0,650 pour (T1), 0,783 pour (T2) et 0,594 pour (T3). Ces résultats montrent clairement que l'usage du langage figuratif complique considérablement l'analyse d'opinions.

ABSTRACT

Figurative language detection has gained relevance recently, due to its importance for efficient sentiment analysis. For the first time, the Défi Fouille de Texte (DEFT) shared task aims to analyse the impact of figuratif language (focusing in particular on irony, sarcasm and humor) on sentiment analysis of French tweets. Three tasks with an increasing level of complexity have been proposed to participants : (T1) polarity analysis of non figurative tweets, (T2) figurative language detection, and (T3) polarity analysis of non figurative and figurative tweets. Twelve teams participated in the competition. Best results in terms of macro f-score are 0.650 for (T1), 0.783 for (T2) and 0.594 for (T3). These results show that the presence of figurative devices make sentiment analysis of tweets much more complex.

MOTS-CLÉS : Analyse d'opinions, langage figuratif, analyse de polarité.

KEYWORDS: Sentiment analysis, figurative language, polarity analysis.

1 Introduction

L'analyse des opinions et sentiments est un domaine de recherche extrêmement actif en traitement automatique des langues, d'autant plus que ces dernières années ont vu se multiplier les sources de données textuelles porteuses d'opinions disponibles sur le web. Devant cette abondance de données et de sources, l'automatisation de la synthèse des multiples avis devient cruciale pour obtenir efficacement une vue d'ensemble des opinions sur un sujet donné. Les systèmes actuels obtiennent de bons résultats sur la classification automatique du caractère subjectif ou objectif d'un document (Liu, 2015). En revanche, les résultats sur la tâche d'analyse de polarité restent encore peu concluants. La raison principale de cet échec est l'incapacité des algorithmes actuels à comprendre toutes les subtilités du langage humain. Parmi ces subtilités, nous nous focalisons sur le langage figuratif.

Contrairement au langage littéral, le langage figuratif détourne le sens propre pour lui conférer un sens dit figuré ou imagé, comme l'ironie, le sarcasme, l'humour, la métaphore ou encore les jeux de mots. La détection automatique du langage figuratif est un sujet de recherche extrêmement actif principalement en raison de son importance pour améliorer les performances des systèmes d'analyse d'opinions (Maynard & Greenwood, 2014; Ghosh *et al.*, 2015; Benamara *et al.*, 2017). Pour ce défi, nous nous intéressons en particulier à l'ironie, au sarcasme et à l'humour.

L'ironie est un phénomène complexe largement étudié en philosophie et en linguistique (Grice *et al.*, 1975; Sperber & Wilson, 1981; Utsumi, 1996). L'ironie est une figure de rhétorique par laquelle on dit le contraire de ce que l'on veut exprimer. Par exemple, pour exprimer une opinion négative envers son téléphone portable, on peut utiliser une forme littérale (« *Ce téléphone est un désastre* ») ou alors une forme imagée (« *Quel super téléphone !* »). En linguistique informatique, l'ironie est un terme générique employé pour désigner un ensemble de phénomènes figuratifs incluant le sarcasme, même si ce dernier s'exprime avec plus d'aigreur et d'agressivité (Clift, 1999).

La détection du langage figuratif et son rôle dans l'analyse de sentiments a fait l'objet de plusieurs campagnes d'évaluation ces dernières années, telles que la campagne SemEval 2015 Task 11 (Ghosh *et al.*, 2015) sur des tweets en anglais et les campagnes SENTIPOLC@Evalita dans leurs éditions de 2014 et 2016 sur des tweets en italien (Basile *et al.*, 2014; Barbieri *et al.*, 2016). Cette édition DEFT est la première campagne d'évaluation autour de ces thèmes pour le français.

2 Données

Constitution Le corpus est constitué de tweets en français qui portent sur des sujets d'actualité (politique, sport, cinéma, émissions TV, artistes, etc.) collectés entre 2014 et 2016, en fonction de la présence de mots-clés (*Hollande, Valls, #DSK, #FIFA,...*) et/ou de hashtags spécifiques indicateurs du langage figuratif (*#ironie, #sarcasme, #humour, #joke*) (Karoui *et al.*, 2017).

Pré-traitements Nous avons supprimé ces hashtags révélant le langage figuratif, aussi bien pour la phase d'annotation manuelle (pour ne pas influencer l'annotateur) que lors de la campagne d'évaluation (pour éviter de fournir un indice fort aux participants). Nous avons également supprimé les retweets, les doublons et les tweets contenant des images [1]. Les emojis et les symboles ont été remplacés par leur code Unicode dans le texte des tweets.

1. Les images ont besoin d'être interprétées pour pouvoir comprendre le langage figuratif du texte associé.

Enfin, il a été recommandé à chaque participant d'utiliser exclusivement le texte des tweets fourni par les organisateurs et non le tweet publié sur Twitter. Ainsi, les identifiants des tweets d'origine ont été remplacés par des identifiants internes à la compétition.

Au total, 7 724 tweets ont été collectés.

3 Tâches proposées

Nous avons proposé trois tâches d'analyse des tweets centrées sur l'analyse d'opinion et du langage figuratif et de niveaux de complexité croissants. Les participants pouvaient choisir de s'inscrire à une ou plusieurs tâches.

Tâche 1. Classification des tweets non figuratifs selon leur polarité.
Étant donné un tweet n'utilisant pas de langage figuratif, cette tâche consiste à le classer selon l'opinion/sentiment/émotion exprimé par son auteur, en : *objectif, positif, négatif* ou *mixte*.

Tâche 2. Identification du langage figuratif.
Étant donné un tweet, cette tâche consiste à identifier s'il contient du langage figuratif (ironie, sarcasme, humour) ou non.

Tâche 3. Classification des tweets figuratifs et non figuratifs selon leur polarité.
Étant donné un tweet utilisant du langage figuratif (ironie ou sarcasme, en excluant les tweets humoristiques) ou non, cette tâche consiste à le classer selon l'opinion/sentiment/émotion exprimé par son auteur en : *objectif, positif, négatif* ou *mixte*.

4 Annotation du corpus

Tous les tweets de notre corpus ont été annotés à la fois en figuratif/non figuratif (cf. section 4.1.1) et en polarité (cf. section 4.1.2) en suivant le guide d'annotation présenté ci-dessous. Nous donnons également dans cette section les accords inter-annotateurs obtenus et la répartition des tweets pour chaque tâche.

4.1 Guide d'annotation

4.1.1 Annotation des tweets selon l'usage ou non du langage figuratif

Pour l'annotation du langage figuratif, il s'agit d'identifier si un tweet utilise du langage figuratif ou non. On s'intéresse à trois phénomènes :

— **L'humour** : ce dernier peut se manifester par l'usage de jeux de mots (exemple 1), de blagues (exemple 2) ou de parodies (exemple 3).

(1) *Si Morandini meurt subitement dans son émission "vous êtes en direct" on pourra dire qu'il est morandirect ? #MDRRRR*

(2) *L'Angleterre le seul pays qui quitte deux fois l'euro en 4 jours #Brexit #Angleterre #ANGISL*

(3) *#Remaniement Franck Ribéry : «Je n'ai jamais pensé être ministre de l'éducation » [Interview] - https ://edukactus.wordpress.com/2014/03/27/*

— **L'ironie et le sarcasme**, comme le montrent les tweets (4), (5) et (6) ci-dessous :

(4) *Ca va bien en Corée du Nord, ils ouvrent un super parc aquatique ! #northkorea*

(5) *Les arbitres étaient digne de la ligue 1 de football tellement ils étaient bons*

(6) *J'adore le taff , manger en 5 minutes et travailler jusqu'a 20h c'est top*

Si un tweet contient au moins une expression relevant de ces phénomènes, il est considéré comme figuratif, sinon il est non figuratif, comme c'est le cas pour les tweets (7) et (8) :

(7) *C'est dommage, Émilie avait l'air bien #adp #ADP2016*

(8) *Le régime de Bachar al-Assad regagne du terrain en #Syrie*

4.1.2 Annotation des tweets selon la polarité

Tous les tweets sont également annotés selon la polarité (à l'exception des tweets humoristiques qui n'expriment pas forcément une opinion). Nous nous intéressons à l'opinion/sentiment/émotion exprimé par l'auteur du tweet envers un sujet. Le tweet peut contenir plusieurs opinions et également plusieurs sujets. L'opinion peut être explicite (avec l'utilisation de mots d'opinion explicitement positifs ou négatifs) ou implicite. La polarité globale du tweet peut prendre une valeur parmi les classes mutuellement exclusives : *objectif, positif, négatif* ou *mixte*.

Objectif : l'auteur du tweet relate des faits ou événements d'actualité, cite une déclaration ou un extrait (titre de journal, etc.) sans donner son opinion. La citation peut contenir une opinion mais nous avons estimé qu'elle était l'opinion de la personne citée et non obligatoirement l'opinion personnelle de l'auteur du tweet.

Exemples de tweets non figuratifs objectifs :

(9) *Syrie : Le Pentagone affirme avoir tué le chef de Khorasan - Monde - lematin.ch*

(10) *#Poutine critique les Etats-Unis sur la question des frappes en #Syrie #SaveSyria*

(11) *Cécile Duflot : "je ne crois pas que DSK soit en mesure de donner des leçons"*

(12) *Je reprends le replay de #DALS*

Dans le cas de tweets figuratifs objectifs (exemple 13), l'auteur relate un ou des faits dont la situation est ironique (ironie situationnelle, ironie du sort, etc.) ou exprime des faits/événements objectifs qui peuvent être ironiques (citation ironique, parodie ironique, titres de journaux ironiques, etc.).

(13) *#DSK songe à déménager en Ontario s'il est libéré des accusations de proxénétisme... http ://www.ledevoir.com/societe/justice/345944/maisons-closes-la-cour-d-appel-de-l-ontario-invalide-la-loi-federale*

Positif : l'auteur du tweet exprime une opinion personnelle uniquement positive sur des faits, des événements ou une citation.

— Exemples de tweets non figuratifs positifs :

(14) *A voir ce soir sur Fr3 l'excellent doc sur l'affaire du #Carlton #dsk*

(15) *Ce soir la reprise du #MeilleurPatissier @M6 @LMP_M6 je dis oui*

— Exemples de tweets figuratifs positifs :

(16) *@lesinrocks La pollution n'est pas une si mauvaise chose : gratuité des transports et journalistes inspirés http ://ow.ly/uClzW*

Négatif : l'auteur du tweet exprime une opinion personnelle uniquement négative sur des faits, des événements ou une citation.

— Exemples de tweets non figuratifs négatifs :

(17) *"La Russie ne va pas imposer la paix à coup de bombes" Ok Obama l'hôpital la charité tout ça.*

(18) *#Valls n'est pas le super héros qu'il s'imagine être*

— Exemples de tweets figuratifs négatifs :

(19) *Bonne nouvelle ! #chômage "@lemondefr : Alerte : le taux de chômage atteint son plus haut depuis juillet 1997*

(20) *J'adore le taff , manger en 5 minutes et travailler jusqu'a 20h c'est top*

Mixte : l'opinion de l'auteur n'est pas exprimée explicitement, elle est sous-entendue ou bien mitigée, à la fois positive et négative.

— Exemples de tweets non figuratifs mixtes :

(21) *Fillon qui souhaite bonne chance à Valls. Je sais pas si je pleure ou si je ris. J'hésite.*

(22) *J'ai faillit pleurer devant la fin de fast and furious 7*

(23) *Et #Sarkozy qui veut nous refaire le coup du nouveau traité comme en 2005*

— Exemples de tweets figuratifs mixtes :

(24) *Un truc avec DSK, mais quoi ? Aucun site internet n'en parle. Surement parce qu' on ne sait rien de ce qu'il s'est réellement passé ?*

(25) *#Sarkozy contre le droit du sol ? Modifier le nom d'un parti vous change un homme !*

4.2 Procédure d'annotation

Pendant la campagne d'annotation, les 4 annotateurs se sont appuyés exclusivement sur le texte du tweet sans pouvoir recourir à des informations contextuelles supplémentaires telles que le profil de l'auteur du tweet, les retweets, le fil de conversation, les hashtags indicateurs de langage figuratif (*#ironie, #humour*), etc. Les annotateurs ont pu cependant utiliser les URLs éventuellement présentes dans le tweet pour le contextualiser (les pages web pointées par les URLs peuvent avoir disparu depuis la collecte).

Les accords inter-annotateurs avant adjudication varient de $\kappa = 0,69$ pour l'annotation du langage figuratif à $\kappa = 0,82$ pour l'annotation de la polarité. Une phase d'adjudication entre 2 annotateurs a été réalisée pour tous les tweets où existait un désaccord. Tous les tweets où un désaccord subsistait après la phase d'adjudication ont été supprimés, soit 407 tweets. Le corpus final se compose de 7 317 tweets.

4.3 Statistiques

Le tableau 1 présente la distribution des 7 317 tweets annotés pour chaque catégorie. Nous observons que les tweets ironiques sont très majoritairement de polarité négative. En effet, l'ironie est souvent employée pour critiquer ou exprimer un mécontentement.

	Non figuratif	Figuratif		TOTAL
		Ironie-Sarcasme	*Humour*	
Objectif	2 053	94	-	**2 147**
Positif	617	12	-	**629**
Négatif	1 585	1 245	-	**2 830**
Mixte	627	166	-	**793**
TOTAL	**4 882**	**1 517**	**918**	**7 317**

TABLE 1 – Distribution des tweets par catégorie.

Les données d'entrainement et de test pour chaque tâche sont réparties comme suit :

— les données de la tâche 1 sont constituées de 4 882 tweets non figuratifs
— les données de la tâche 2 sont constituées des tweets de la tâche 1 et de 2 435 tweets figuratifs
— les données de la tâche 3 sont les tweets de la tâche 2 pour lesquels nous avons retiré les 918 tweets humoristiques

La répartition entre les données d'entrainement et de test est respectivement de 80 % et 20 %, en conservant les mêmes proportions pour chaque classe. Les tableaux 2, 3 et 4 montrent cette répartition pour les trois tâches.

Tâche 1	Entrainement	Test	TOTAL
Objectif	1 642 (42,1%)	411 (42%)	**2 053 (42,05%)**
Positif	494 (12,6%)	123 (12,65%)	**617 (12,65%)**
Négatif	1 268 (32,5%)	317 (32,5%)	**1 585 (32,45%)**
Mixte	502 (12,8%)	125 (12,85%)	**627 (12,85%)**
TOTAL	**3 906 (80%)**	**976 (20%)**	**4 882**

TABLE 2 – Répartition des tweets pour la tâche 1

5 Organisation de la compétition

La compétition s'est déroulée en deux temps. D'abord une phase d'inscription, allant du 3 février 2017 au 1er mai 2017 pendant laquelle les participants ont reçu les données d'entraînement. **18**

Tâche 2	Entrainement	Test	TOTAL
Non figuratif	3 906 (66,7 %)	976 (66,7 %)	**4 882 (66,7 %)**
Figuratif	1 947 (33,3 %)	488 (33,3 %)	**2 435 (33,3 %)**
TOTAL	**5 853 (80 %)**	**1 464 (20 %)**	**7 317**

TABLE 3 – Répartition des tweets pour la tâche 2

Tâche3	Entrainement			Test			TOTAL
	Non figuratif	*Figuratif*	Total	*Non figuratif*	*Figuratif*	Total	
Objectif	1 642 (42,1 %)	75 (6,2 %)	1 717 (33,55 %)	411 (42 %)	19 (6,23 %)	430 (33,57 %)	**2 147 (33,55 %)**
Positif	494 (12,6 %)	10 (0,8 %)	504 (9,85 %)	123 (12,65 %)	2 (0,66 %)	125 (9,76 %)	**629 (9,83 %)**
Négatif	1 268 (32,5 %)	995 (82,1 %)	2 263 (44,2 %)	317 (32,5 %)	250 (81,97 %)	567 (44,26 %)	**2 830 (44,22 %)**
Mixte	502 (12,8 %)	132 (10,9 %)	634 (12,4 %)	125 (12,85 %)	34 (11,14 %)	159 (12,41 %)	**793 (12,4 %)**
Total	3 906 (76,3 %)	1 212 (23,7 %)	5 118	976 (76,2 %)	305 (23,8 %)	1 281	**6 399**
TOTAL	**5 118 (80 %)**			**1 281 (20 %)**			**6 399**

TABLE 4 – Répartition des tweets pour la tâche 3

équipes se sont inscrites. Ensuite, la phase de test s'est déroulée du 2 au 9 mai 2017 dans une période de trois jours au libre choix de chaque équipe (accès aux données de test le premier jour, soumission des fichiers de résultats avant la fin du troisième jour). Il n'y avait pas de limite quant au nombre de tâches auxquelles pouvait participer une équipe. Le nombre maximum de systèmes différents présentés par une équipe pour une tâche donnée a été cependant limité à 3. Finalement, **13** équipes ont participé dont **6** équipes académiques françaises, **1** équipe issue de l'industrie, **2** équipes mixtes académique/industrie et **4** équipes académiques venant d'Inde, du Canada, du Maroc, et une équipe mixte venant de Belgique et de Turquie. Pour la tâche 1, nous avons reçu un total de 33 soumissions correspondant à l'ensemble des 12 équipes ayant participé au défi. Sur la tâche 2, nous avons reçu 31 soumissions pour 11 équipes, et 24 soumissions pour 9 équipes sur la tâche 3.

L'atelier de clôture s'est tenu le 26 juin, pendant la conférence TALN/RECITAL 2017 à Orléans.

6 Méthodes des participants

Il est intéressant de constater que la plupart des participants n'ont pas eu recours à des méthodes spécifiques pour la détection du langage figuratif. En effet, les mêmes approches ont été utilisées quelle que soit la tâche. Tous les participants ont en revanche utilisé des méthodes par apprentissage supervisé : réseaux de neurones, SVM, classifieur bayésien naïf, K plus proches voisins, boosting d'arbres de décision (l'équipe 12 LIUM-OCTO a utilisé une combinaison d'un système à base de règles et de classifieurs).

Ces modèles utilisent entre autres :

— des lexiques dédiés : par exemple, des lexiques de sentiments et d'émotions comme FEEL, Affect, Polarimots, Diko et labMT (équipe 2 Tweetaneuse, équipe 4 Advanse-LIRMM, équipe 12 LIUM-OCTO, équipe 16 OrangeLabs), NRC-EmoLex traduit en français (équipe 1 TW-StAR).

— des représentations en sac de mots : représentations tf-idf de bigrammes (équipe 14 Melodi-IRIT), présence/absence d'unigrammes, bigrammes et trigrammes (équipe 4 Advanse-LIRMM), vecteurs d'unigrammes (équipe 15 Amrita University).

— des représentations en plongements de mots (*word embeddings*). Des modèles vectoriels obtenus avec Doc2Vec, Word2Vec, Fasttext ou encore SkipGram ont été utilisés (équipe 17 LS2N, équipe 2 Tweetaneuse, équipe 14 Melodi-IRIT, équipe 5 IRISA). Afin d'intégrer la polarité des mots dans la représentation vectorielle, l'équipe 8 LIA a proposé en plus des représentations à base de *sentiment embedding* qui ont permis d'entraîner des réseaux de neurones convolutionnels.

En plus des méthodes listées plus haut, l'équipe 2 Tweetaneuse a proposé une approche fondée sur une extraction de motifs en caractères fermés et fréquents combinée avec des algorithmes d'apprentissage automatique. Cette méthode, assez proche des méthodes de stylométrie utilisées pour les tâches d'attribution d'auteur, semble être particulièrement prometteuse pour la tâche 2 de détection du langage figuratif.

7 Évaluation et résultats

7.1 Métriques d'évaluation

Nous avons évalué les soumissions des participants selon les mesures classiques de précision, rappel et f-mesure. Nous avons retenu la macro f-mesure (formule 26) comme mesure officielle car elle met à égalité chacune des classes à traiter, indépendamment de la distribution des données par classe (Manning & Schütze, 2000). Précisons que ce choix défavorise les systèmes qui ne traitent pas les classes minoritaires (par choix ou par manque de robustesse du système). Pour cette raison, nous présentons également le classement obtenu au moyen de la micro-précision (formule 27).

$$\text{Macro F-mesure} = \frac{\sum_{i=1}^{n} \left(\frac{2 \times \text{vrais positifs}(i)}{2 \times \text{vrais positifs}(i) + \text{faux négatifs}(i) + \text{faux positifs}(i)} \right)}{n} \tag{26}$$

$$\text{Micro-précision} = \frac{\sum_{i=1}^{n} \text{vrais positifs}(i)}{\sum_{i=1}^{n} \text{vrais positifs}(i) + \sum_{i=1}^{n} \text{faux positifs}(i)} \tag{27}$$

7.2 Résultats sur l'analyse de polarité (tâche 1 et tâche 3)

Le tableau 5 présente les résultats et le classement des systèmes selon la micro-précision et la macro f-mesure pour la tâche 1. Les meilleurs résultats ont été obtenus par le LIA avec une macro f-mesure de 0,650, suivie de l'équipe Advanse-LIRMM avec 0,557 et de l'équipe MELODI-IRIT avec 0,547. Une analyse détaillée des résultats par classe pour cette tâche revèle que la classe *objective* (42,11% des instances) obtient les meilleurs scores pour tous les participants avec une f-mesure allant de 0,811 à 0,410. On note aussi que, bien que la classe *négative* soit plus fréquente comparée à la classe *positive*, les résultats obtenus par les systèmes des participants pour ces deux classes sont globalement comparables (0, 1 d'écart en moyenne) [2]. Enfin, pour la classe *mixte*, les résultats sont les moins bons (inférieurs à 0,4 en f-mesure) alors que cette classe est aussi fréquente que la classe *positive* (12,81% et 12,6% des instances respectivement). Les tendances sont quasiment les mêmes si on compare les résultats par classe en terme de précision.

Tâche 1	Micro-précision				Macro f-mesure			
	1	2	3	Rang	1	2	3	**Rang**
LIA (équipe 8)	0,682	0,704	**0,710**	1	0,602	0,634	**0,650**	**1**
Advanse, LIRMM (équipe 4)	0,640	0,628	**0,653**	4	0,555	0,539	**0,557**	**2**
MELODI, IRIT (équipe 14)	**0,697**	0,695	0,695	2	**0,547**	0,545	0,546	**3**
LIUM-OCTO (équipe 12)	0,621	**0,625**	0,589	7	0,534	0,539	**0,545**	**4**
LS2N (équipe 17)	0,632	**0,635**	0,634	5	0,493	**0,534**	0,487	**5**
Tweetaneuse (équipe 2)	**0,622**	0,573	0,571	6	**0,531**	0,488	0,498	**6**
IRISA (équipe 5)	0,644	0,659	**0,676**	3	0,512	0,512	**0,514**	**7**
OrangeLabs (équipe 16)	**0,620**			8	0,467	**0,502**		**8**
Tw-StAR (Université Libre de Bruxelles, Selcuk University, Ibn Zohr University ; équipe 1)	0,537	0,230	0,579	9	0,406	0,217	**0,492**	**9**
LIRMM (équipe 18)	0,341	**0,541**		10	0,219	**0,353**		**10**
Amrita University (équipe 15)	**0,387**	0,364	0,351	11	**0,276**	0,228	0,210	**11**
Équipe 13	**0,308**	0,308		12	**0,239**	0,239		**12**

TABLE 5 – Résultats et classement des systèmes selon la micro-précision et la macro f-mesure pour la tâche 1. Le classement est fait sur la base de la meilleure soumission de chaque participant

Lorsque des tweets figuratifs sont introduits à la tâche de détection de polarité (Tâche 3), on observe que les résultats sont nettement moins bons. Ces derniers sont présentés dans le tableau 6 ainsi que le classement des systèmes selon la micro-précision et la macro f-mesure. Les équipes du trio de tête obtiennent respectivement une macro f-mesure de 0,594 (LIA), 0,534 (LS2N) et 0,531 (Advanse-LIRMM). Comme pour la tâche 1, la classe *objective* (33,57% des instances) obtient les meilleurs scores à la fois en terme de macro f-score et précision. Les scores sont très faibles pour la classe *mixte* (moins de 0,31 de f-mesure) alors qu'elle est plus fréquente que la classe *positive* (12,41% et 9,76% des instances respectivement). Enfin, les scores pour les classes *objective* et *negative* sont proches (alors que l'écart était plus grand dans la tâche 1). En effet, la classe *négative* était majoritaire pour cette dernière tâche (44,26% des instances), ce qui est très caractéristique de la polarité des tweets ironiques.

2. A l'exception de l'équipe LIUM-OCTO qui obtient de meilleurs résultats pour la classe *positive*.

Ces résultats montrent que l'analyse de polarité devient plus complexe lorsqu'un renversement de polarité dû à l'usage de l'ironie et du sarcasme est employé. Voici les principales conclusions que l'on peut tirer de la comparaison des résultats de la tâche 3 par rapport à ceux de la tâche 1 :

— Toutes les équipes ont observé une baisse de performance en terme de macro f-mesure, à l'exception de l'équipe 5 (IRISA).
— La plus forte baisse concerne la classe *objective* avec une diminution de la macro f-mesure de 0,071 en moyenne. Ainsi, des tweets ironiques positifs ou négatifs sont souvent classés en objectif. La détection du langage figuratif est donc une étape cruciale pour l'analyse d'opinion.
— Par rapport à la tâche 1 (sans figuratif), tous les participants font mieux sur la classe *négative*.
— Les scores sur les classes *positive* et *mixte* restent stables entre les tâches 1 et 3 car il y a très peu d'instances figuratif-positif et figuratif-mixte dans la tâche 3.

Tâche 3	Micro-précision				Macro f-mesure			
	1	2	3	Rang	1	2	3	**Rang**
LIA (équipe 8)	0,672	0,675	**0,678**	3	0,578	0,585	**0,594**	**1**
LS2N (équipe 17)	**0,649**	0,636	0,639	4	**0,534**	0,471	0,477	**2**
Advanse, LIRMM (équipe 4)	0,634	0,620	**0,639**	6	**0,531**	0,515	0,523	**3**
MELODI, IRIT (équipe 14)	0,681	**0,683**	0,677	2	0,515	**0,522**	0,514	**4**
Tweetaneuse (équipe 2)	**0,627**	0,524	0,546	7	**0,519**	0,482	0,477	**5**
IRISA (équipe 5)	0,643	0,652	**0,688**	1	0,509	0,509	**0,517**	**6**
OrangeLabs (équipe 16)	**0,639**			5	**0,469**			**7**
LIRMM (équipe 18)	0,379	**0,545**		8	0,215	**0,341**		**8**
Amrita University (équipe 15)	0,424	**0,424**	0,422	9	0,220	**0,232**	0,231	**9**

TABLE 6 – Résultats et classement des systèmes selon la micro-précision et la macro f-mesure pour la tâche 3. Le classement est fait sur la base de la meilleure soumission de chaque participant

7.3 Résultats sur la détection du langage figuratif (tâche 2)

Le tableau 7 présente les résultats et le classement des systèmes selon la micro-précision et la macro f-mesure pour la tâche 2. Les meilleurs résultats en terme de macro f-mesure sont 0,783 (LIA), 0,774 (LIUM-OCTO) et 0,750 (Advanse-LIRMM). Pour tous les participants, la classe *non figuratif*, majoritaire (66,66%), obtient le meilleur score avec une moyenne de f-mesure pour l'ensemble des participants de 0,798. La classe *figuratif* obtient des scores nettement plus faibles, autour de 0,590 de f-mesure en moyenne.

8 Conclusion

Pour la première fois, l'édition 2017 du Défi Fouille de Texte (DEFT) s'est intéressée à l'influence du langage figuratif (en particulier l'ironie, le sarcasme et l'humour) dans l'analyse d'opinions dans des tweets en français. Trois tâches de niveaux de complexité croissants ont été proposées aux participants : (T1) déterminer la polarité globale des tweets non figuratifs, (T2) déterminer si un tweet

Tâche 2	Micro-précision				Macro f-mesure			
	1	2	3	Rang	1	2	3	**Rang**
LIA (équipe 8)	**0,807**	0,801	0,802	2	**0,783**	0,774	0,744	**1**
LIUM-OCTO (équipe 12)	0,699	**0,810**	0,721	1	0,659	**0,774**	0,677	**2**
Advanse, LIRMM (équipe 4)	0,788	0,788	**0,790**	3	**0,750**	0,749	0,750	**3**
Tweetaneuse (équipe 2)	**0,779**	0,761	0,761	5	**0,746**	0,716	0,716	**4**
IRISA (équipe 5)	0,779	0,742	**0,782**	4	0,741	0,642	**0,745**	**5**
LS2N (équipe 17)	0,745	**0,758**	0,751	7	0,718	**0,720**	0,704	**6**
LIRMM (équipe 18)	0,587	**0,755**	0,712	8	0,457	**0,715**	0,630	**7**
MELODI, IRIT (équipe 14)	**0,768**	0,755	0,766	6	**0,709**	0,692	0,706	**8**
OrangeLabs (équipe 16)	**0,699**			9	**0,663**			**9**
Équipe 11	0,613	**0,692**	0,688	10	**0,577**	0,550	0,542	**10**
Amrita University (équipe 15)	0,488	0,490	**0,497**	11	**0,557**	0,550	0,542	**11**

TABLE 7 – Résultats et classement des systèmes selon la micro-précision et la macro f-mesure pour la tâche 2. Le classement est fait sur la base de la meilleure soumission de chaque participant

contient ou non du langage figuratif, et (T3) déterminer la polarité globale des tweets figuratifs et non figuratifs. Douze équipes ont participé à ce défi.

Il est intéressant de noter que la plupart des participants n'ont pas eu recours à des méthodes spécifiques pour la détection du langage figuratif et ont utilisé les mêmes approches, par apprentissage automatique supervisé ou non supervisé, pour les 3 tâches. Les meilleurs résultats, en macro f-mesure, sont de 0,650 pour (T1), 0,783 pour (T2) et 0,594 pour (T3). Ces résultats montrent clairement que l'usage du langage figuratif complique considérablement l'analyse d'opinions.

Remerciements

Ces travaux ont été partiellement financés par le projet STAC (ERC 269427).

Références

BARBIERI F., BASILE V., CROCE D., NISSIM M., NOVIELLI N. & PATTI V. (2016). Overview of the Evalita 2016 SENTIment POLarity Classification Task. In *Proc of Third Italian Conference on Computational Linguistics (CLiC-it 2016) and Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, Napoli, Italia : CEUR Workshop Proceedings.

BASILE V., BOLIOLI A., NISSIM M., PATTI V. & ROSSO P. (2014). Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proc of EVALITA*, p. 50–57, Pisa, Italy : Pisa University Press.

BENAMARA F., TABOADA M. & MATHIEU Y. (2017). Evaluative language beyond bags of words : Linguistic insights and computational applications. *Computational Linguistics*, **43**(1), 201–264.

CLIFT R. (1999). Irony in conversation. *Language in Society*, **28**, 523–553.

GHOSH A., LI G., VEALE T., ROSSO P., SHUTOVA E., BARNDEN J. & REYES A. (2015). Semeval-2015 task 11 : Sentiment analysis of figurative language in twitter. In *Proc of SemEval*, p. 470–478, Denver, CO.

GRICE H. P., COLE P. & MORGAN J. L. (1975). Syntax and semantics. *Logic and conversation*, **3**, 41–58.

KAROUI J., BENAMARA F., MORICEAU V., PATTI V., BOSCO C. & AUSSENAC-GILLES N. (2017). Exploring the impact of pragmatic phenomena on irony detection in tweets : A multilingual corpus study. In *Proc of EACL*, Valencia, Spain.

LIU B. (2015). *Sentiment Analysis : Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

MANNING C. D. & SCHÜTZE H. (2000). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts : MIT Press.

MAYNARD D. & GREENWOOD M. A. (2014). Who cares about sarcastic tweets ? investigating the impact of sarcasm on sentiment analysis. In *Proc of LREC*, p. 4238–4243, Reykjavik, Iceland.

SPERBER D. & WILSON D. (1981). Irony and the use-mention distinction. *Radical pragmatics*, **49**, 295–318.

UTSUMI A. (1996). A unified theory of irony and its computational formalization. In *Proc of COLING*, p. 962–967, Copenhagen, Denmark.